INTERSPEECH 2009 BRIGHTON

On the Cost of Backward Compatibility for Communication Codecs

Konstantin Schmidt[1], *Markus Schnell*[1], *Nikolaus Rettelbach*[1], *Manfred Lutzky*[1], *Jochen Issing*[1]

[1]Fraunhofer Institute for Integrated Circuits, Erlangen, Germany

Konstantin.Schmidt@iis.fraunhofer.de

Abstract

Super wideband (SWB) communication calls more and more attention as can be seen by the standardization activities of SWB extensions for well-established wideband codecs, e.g. G.722 or G.711.1. This paper presents a technical solution for extending the G.722 codec and compares the new technology to other standardized SWB codecs. Hereby, a closer look is given on the concept of extending technologies to more capabilities in contrast to non-backwards compatible solutions.

Index Terms: noise shaping, backward compatible audio coding, G.722, spectral band replication

10.21437/Interspeech.2009-689

1. Motivation

The introduction of new telephony services, e.g. VoIP over broad band networks, enables providers to offer communication products of higher quality. Therefore, many recent standardization activities inside ITU-T SG16 target at the extension of existing codecs towards higher bandwidth and audio quality in a backward compatible manner. One recent example is ITU-T SG16 super wideband extension of G.722 and G.711.1 [1].

Backward compatibility is achieved through adding new enhancement layers to legacy bitstreams in order to maintain the interoperation with legacy devices. New terminals are able to decode the legacy parts together with the additional layers and thus, are able to provide enhanced quality. This kind of backward compatibility is also known as embedded coding (ITU-T) or scalable coding (MPEG). The reasoning for backward compatibility is to enhance existing telephone infrastructure while keeping compatibility to legacy equipment at the same time. On the other hand, this concept implies that outdated technology is carried through the network although much more sophisticated algorithms might be available. The purpose of this paper is to assess this drawback by designing a super wideband extension (section 4) for the widespread G.722 codec (section 3.1) and to compare this technology to other state-of-the-art communication codecs (section 3.2) in terms of quality (section 5) and interoperability (section 2).

2. Backward Compatibility

The main purpose for backward compatibility is to seemlessly preserve usability with technology upgrades. Anyhow, backward compatibility can be achieved in different ways:

- **backward compatible signalling**, where all devices share one common base codec, which is used for communication with legacy devices. New devices may also provide high quality codecs for communication. Through codec negotiation, e.g. by Session Initialization Protocol (SIP), the best common codec of all involved devices is chosen and thus, backwards interoperability is assured.
- **backward compatible codec**, where a legacy equipped device can decode a bit stream produced by a new device using a new codec and vice versa. This bit stream consists of a legacy base layer and additional enhancement layers whereas the enhancement layers code the upper part of the spectrum or extra channels.
- **scalable codec**, where a codec contains of a core codec (often a speech codec) and enhancement layers that code the core codec's error signal [2], [3]. The enhancement layers can be of the same or different type than the core codec.

The last two concepts of backward compatible codecs can offer benefits considering the following scenarios:

- **inter network/terminal handover** In case an established call is handed over due to abandonment of coverage range from one to another network with different codec capability, e.g. from upcoming Evolved Packet Service codec over LTE (3GPP Long Term Evolution) to Adaptive Multi-Rate Wideband (AMR-WB) [4] over 3GPP, a re-negotiation of codec capability can be avoided and the call transfer is speeded up. The same behaviour can be observed for a call transferred within the same network from one terminal to another with different codec capability.
- **three party conference without Multipoint Control Unit** In case a three party peer to peer conference is established without using a Multipoint Control Unit that handles the mixing of individual speech signals, every participant sends a dedicated bitstream to every other participant. Even if participants have different coding capability only a single encoder instance can be deployed and so workload is saved and battery lifetime increased.

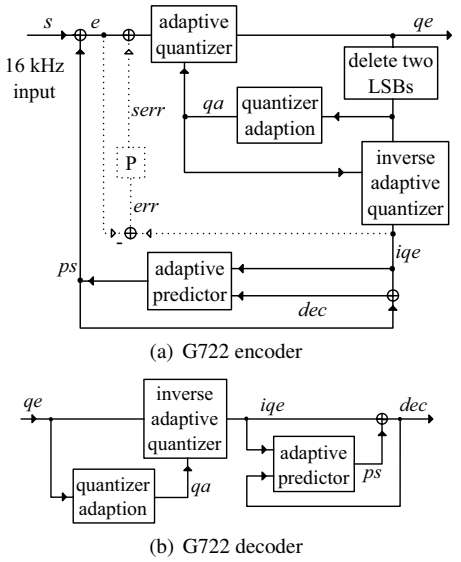
On the other hand, the method of backward compatible signalling allows the freedom to choose a specific codec offering better efficiency or more features. This degree of freedom has to be balanced against the mentioned advantages of backward compatible codecs.

3. Technology Under Consideration

In this section we will give a short overview of ITU-T G.722 speech codec, which we decided to base our backward compatible codec design on. Furthermore we will illustrate MPEG-4 Enhanced Low Delay AAC (AAC-ELD) as an example for a non backward compatible high band width communication audio codec.

3.1. G.722

The core codec used in the proposed system is the well-known ITU-T codec G.722 which has already been standardized in



(a) G722 encoder

(b) G722 decoder

Figure 1: Structure of G.722 codec including noise shapeing explained in section 4.2

1988 [5]. The codec is based on subband adaptive differential pulse code modulation (SB-ADPCM) and is designed for a bandwidth of 50-7000 Hz, thus it is a wideband codec. Figure 1 shows the encoder and the decoder at the lower band including - drawn in dotted lines - the noise shaping filter (explained in section 4.2) which is not part of the standard. The input signal is sampled at 16 kHz and then split into two subbands (0-4 kHz and 4-8kHz) using a linear-phase quadrature modulated filterbank (QMF). Subsequently each subband is downsampled to 8 kHz and encoded via ADPCM. The higher subband is always encoded with 2 bit per sample while the lower subband can be encoded with 4 - 6 bits per sample yielding a total bitrate of 48 - 64 kbit/s.

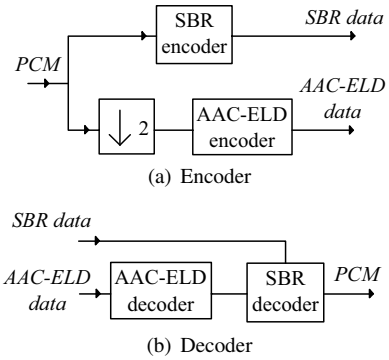
The basic idea behind ADPCM is to code the difference between the input sample and an estimation of this sample predicted with some sort of adaptive prediction algorithm. In G.722 the predictor is adapted sample wise on past samples thus the ADPCM algorithm used here is delay free (neglecting the QMF). The recursive least mean square (LMS) algorithm that is used to adapt the predictor coefficients is described in detail in [6].

The adaptive predictor used in G.722 comprises a second order section that models poles and a sixth order section that models zeros in the input signal [7]. The zero-section is adapted on the inverse quantized signal iqe while the pole-section is adapted on the sum dec of the inverse quantized signal iqe and the predicted signal ps. Comparing the structure of the encoder with the structure of the decoder one can see that - in case the codewords are transmitted without error - both signals iqe and dec at the encoder are identical to the signals iqe and dec at the decoder. Thus in G.722 no side info has to be transmitted besides the quantized prediction error.

The quantizer is a memoryless logarithmic quantizer which is scaled/adapted to the signals variance. Further details are described in [7].

3.2. MPEG-4 AAC-ELD

Standardized recently in ISO/IEC 14496-3 [8], MPEG-4 Enhanced Low Delay AAC (AAC-ELD) [9] represents the combination of the MPEG-4 Low Delay AAC (AAC-LD) with the MPEG-4 Spectral Band Replication (SBR) tool. The latter one serves in the list of MPEG tools as a generic extension coder for higher frequencies, especially for very low bit rate coding. The concept of SBR allows the encoding of the lower frequencies by the transformation based AAC core coder while the upper frequencies can be regenerated from the core signal with the help of some additional guidance information. Figure 2 gives an overview of the internal structure of AAC-ELD.



(a) Encoder

(b) Decoder

Figure 2: Structure of AAC-ELD

The full signal is analyzed by the SBR encoder where the guidance information for the regeneration process is extracted. In parallel, the signal is downsampled and encoded by the AAC-ELD core coder employing the main coding tools of AAC, as described in [10]. At the decoder side, the core coder extracts the lower frequency part from the bit stream and passes the audio data on to the SBR decoder where the signal is upsampled and reconstructed.

The lower frequency parts are transposed up to the higher regions and further spectral adjustments are applied. These consist of a *level* and *noise floor* adjustment and further on, a *tonality correction* which allows the insertion of missing harmonic components or the erasure of them. Due to the fact, SBR is operating in a filter bank with a high temporal resolution, it allows for an accurate reconstruction of the temporal shape of the high band signal as well. More detailed descriptions of SBR can be found in [11].

The main innovation of AAC-ELD compared to its precursor AAC-LD and SBR is the employment of low delay filter banks [12]. By introducing this technology to the two modules, their combination become possible in the first place by maintaining the algorithmic delay low enough to be suitable for bi-directional communications. The codec offers an overall algorithmic delay of around 32 ms if the SBR tool is used, as described above, and a delay of 15 ms only if the AAC-ELD core coder works in stand-alone mode.

4. Super Wideband Extension for G.722

4.1. System design

Considering the current super wideband standardization activities in ITU-T SG16 we decided to base our backward compatible system design on ITU-T G.722 and enhance it by bitstream

compatible changes to the core encoder and extend its audio band width by applying a state of the art band width extension (BWE). As a natural BWE candidate we decided for the SBR tool as used in AAC-ELD, as it is already optimized for low delay and delivers a high audio quality. Figure 3 pictures an overview of the proposed system.
The input signal is assumed to be sampled at 32 kHz sampling frequency. The SBR module extracts the SBR parameters as described in section 3.2 on blocks of 5 ms. The built in quadrature modulated filterbank (QMF) of the SBR module downsamples the signal to 16 kHz. On this downsampled signal linear predictive coefficients (LPC) are calculated to be used for shaping the quantization noise as explained further in section 4.2. The linear predictive coefficients are calculated every 5 ms by applying the Levinson-Durbin recursion on autocorrelation values calculated on windowed blocks of 10 ms length. The window function is the same asymmetric sinusoidal window as described in [1]. The delay of the codec is caused by the blocking for calculating the LPCs and the QMF. The block length is - as already mentioned - 5 ms but the QMF needs 1 ms lookahead and thus the total codec delay results in 6 ms.

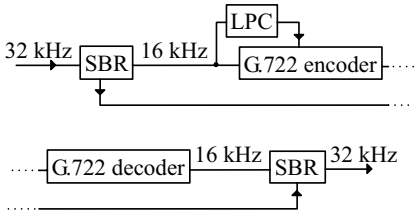


Figure 3: Overview of the proposed codec (encoder top, decoder below)

The bitstream is composed of 56 kbit/s quantized prediction error from the G.722 core coder and 8 kbit/s SBR parameters. The SBR parameters consist of the *noise floor*, the *tonality* and the *QMF energy*. Since the frame size is 5 ms at a sampling frequency of 32 kHz there are 40 bits per frame to be used for coding the SBR parameters. The first QMF energy is quantized with a logarithmic Lloyd-Max [13] quantizer and coded absolute, while for the remaining QMF energies the difference to the first energy is encoded, again with a Lloyd-Max quantizer. The noise floor as well as the tonality values is quantized linear and coded separately.

4.2. Noise Shaping

It can be shown that in closed-loop ADPCM - unlike in open loop ADPCM - the coding error (the difference between the input signal and the decoded output signal) is identical to the quantization noise from the quantizer [14]. In G.722 the error signal generated by the quantizer is white, thus it has a flat spectrum. In audio coding however, the quantization noise should follow the masking threshold to minimize the audible distortion.

Fortunately the structure of the ADPCM codec allows for shaping the spectral characteristics of the quantization noise signal adaptive by applying *noise shaping* [15] and - furthermore - this can be done without needing to change the decoder. By applying noise shaping the spectral characteristic of a quantizer can be adapted to the masking curve provided that there is access to the quantization error. As shown in figure 1(a) the quantization error err is calculated as difference between the quantizer's input signal and the dequantized signal. This error signal is weighted with some filter coefficients P and then fed back to the quantizer's input signal. This is an infinite impulse response (IIR) filter system; once $err \neq 0$, the loop around the quantizer will continue indefinitely, even if the quantization noise would suddenly vanish. The transfer function of the filter P affecting the quantization noise is given by:

$$P(z) = \frac{1}{1 - \sum_{i=1}^{12} \gamma_i a_i z^{-i}}, \quad (1)$$

where γ is a weighting factor to guarantee the filter stability. It is usually set to 0.92 but changed for signals with spectral tilt towards high-frequencies and signals with low energy (for further explanations see [16]).

As already shown in [16] using linear prediction coefficients (LPC - here calculated according to figure 3) as filter coefficients a_i in P will lead to a PSD of the error signal following a rough approximation of the masking threshold. In the proposed system a 12th order filter is used for shaping the error signal. This order is chosen as a tradeoff between quality and complexity. A higher filter order allows to "shape" the quantization noise into more spectral peaks, while a lower filter order is less complex to calculate.

5. Evaluation

5.1. Methodology

To evaluate the audio quality of the proposed coding system a listening test was conducted. The set of items for this test has been used in many MPEG audio tests and consists of 12 different items that proved to be critical. The items contain speech (es01, es02, es03), general music (sc01, sc02, sc03) and single instruments (si0x, sm0x) with special characteristics as for example transient signals like castanets (si02) or very tonal items like a pitch pipe (si03). The G.722 SWB codec, as described in Section 4, operates at a bitrate of 64 kbps total. The state of the art communication codecs G.722.1 Annex C [17], G.719 [18] and AAC-ELD are operated at a bit rate of 32 kbps to account for the advantage of non backward compatibility. The listening test was conducted according to MUSHRA [19] test methodology. The codecs under test were presented to the subjects together with two band limited anchors (3.5 and 7 kHz) and the hidden original. The items were played out from a fanless computer equipped with a professional sound card, Stax headphones and amplifier were used. As listening environment, a laboratory was chosen that was not anechoic but dedicated for listening tests. Eight expert listeners participated in the listening test.

5.2. Results and Discussion

Figure 4 shows the mean and 95% confidence intervals of the listening test. AAC-ELD performs significantly better than all other codecs in the test followed by G.722 SWB. G.722 SWB technology shows a significant better performance over all items compared to the G.722.1 Annex C and the G.719 codec. Comparing AAC-ELD with the worst codec in the test G.719 shows a difference of app. 20 Mushra scores, which corresponds roughly to the perceived quality difference between narrow and a wide band audio. Considering that AAC-ELD operates at half the bit rate of G.722 SWB, this experiment demonstrates impressively the conceptual predominance of AAC-ELD over a backward compatible codec design in terms of compression efficiency.

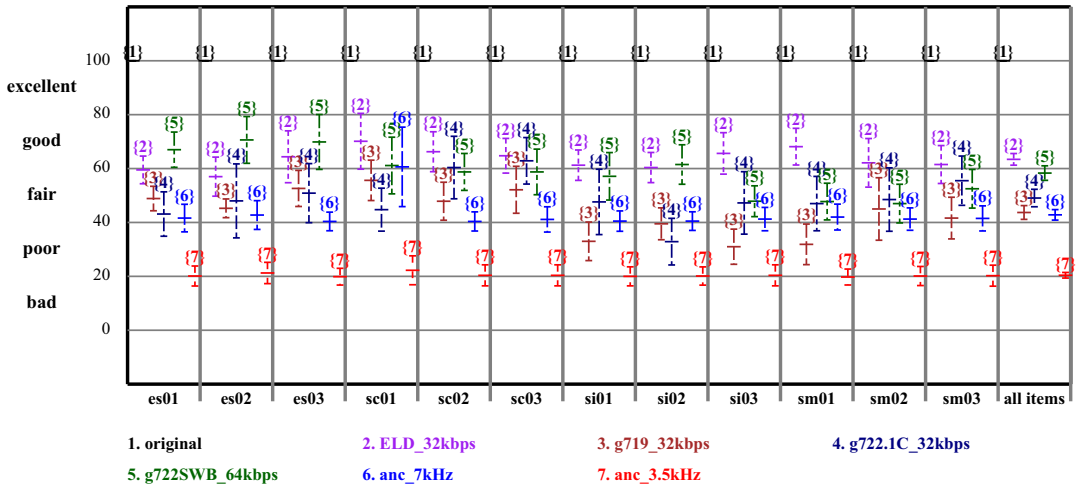


Figure 4: Mean and 95% confidence intervals of MUSHRA listening test

6. Conclusions

We have discussed aspects of backward compatible coding (embedded coding) in telecommunication systems and presented a possible system design for a super wide band extension to the ITU-T G.722 speech coding algorithm. The designed system was compared to state of the art super wide and full band audio coding schemes for communicational systems. Also delivering fair to good sound quality and even outperforming some state of the art codecs in terms of quality, the proposed system has the burden of a core codec design that does not allow for a efficient compression performance. AAC-ELD is able to deliver significantly better audio quality by utilizing only half the bitrate. Given the fact that by using a backward compatible signalling like SIP no drawback in terms of audio quality can be experienced compared to utilisation of backward compatible codecs and that the advantages of enhancing an existing service network through a backward compatible codec occurs in rather limited use cases, but comes at the cost of significantly reduced codec efficiency, we think that new network services should be accomplished by codec designs solely based on state of the art technology.

7. References

[1] Y. Hiwasaki, S. S. ans Hitoshi Ohmuro, T. Mori, J. Seong, M. S. Lee, B. Kövesi, S. Ragot, J.-L. Garcia, C. Marro, L. Miao, J. Xu, V. Malenovsky, J. Lapierre, R. Lefebvre, "G.711.1: A WIDEBAND EXTENSION TO ITU-T G.711", *106th European Signal Processing Conference*, 2008

[2] K. Brandenburg, B. Grill, "First Ideas on Scalable Audio Coding", *97th AES Convention*, San Francisco, USA, preprint 3924, Nov. 1994

[3] B. Grill, "A bit rate scalable perceptual coder for MPEG-4 Audio", *103rd AES Convention*, New York, USA, preprint 4620, Oct. 1997

[4] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, , K. Järvinen, "The Adaptive Multirate Wideband Speech Codec (AMR-WB)", *IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING*, 2002

[5] International Telecommunication Union, "7 kHz audio-coding within 64 kbit/s ", ITU-T Recommendation G.722, 1988

[6] P. L. Feintuch, "An Adaptive Recursive LMS Filter", *Proc. IEEE*, 1976

[7] X. Maitre, "7 kHz Audio Coding Within 64 kbit/s", *IEEE Journal on Selected Areas in Communication*, 1988

[8] ISO/IEC 14496-3:2008, "Coding of Audio-Visual Objects, Part 3: Audio - Amendment 9", 2008

[9] M. Schnell, M. Schmidt, M. Jander, T. Albert, R. Geiger, V. Ruoppila, P. Ekstrand, M. Lutzky, B. Grill, "MPEG-4 Enhanced Low Delay AAC - a new standard for high quality communication", *125th AES Convention*, San Francisco, CA, USA, preprint 7503, Oct. 2008

[10] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, Y. Oikawa, "ISO/IEC MPEG-2 advanced audio coding", *J. Audio Eng. Soc.*, vol. 45, no. 10, pp. 789–813, Oct. 1997

[11] M. Dietz, L. Liljeryd, K. Kjörling, O. Kunz, "Spectral Band Replication, a Novel Approach in Audio Coding", *112th AES Convention*, Munich, Germany, preprint 5553, Apr. 2002

[12] M. Schnell, R. Geiger, M. Schmidt, M. Multrus, M. Mellar, J. Herre, G. Schuller, "Low Delay Filterbanks for Enhanced Low Delay Audio Coding", *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, pp. 235–238, Oct. 2007

[13] S. P. LLOYD, "Least Squares Quantization in PCM", *IEEE TRANSACTIONS ON INFORMATION THEORY*, 1982

[14] M. Holters, C. R. Helmrich, U. Zölzer, "Delay-free audio coding based on ADPCM and error feedback", *11th International Conference on Digital Audio Effects*, 2008

[15] U. Zölzer, *Digitale Audiosignalverarbeitung*, B.G. Teubner, (in German), 1997

[16] J. Lapierre, R. Lefebvre, B. Bessette, V. Malenovsky, R. Salami, "Noise shaping in an ITU-T G.711-Interoperable embedded codec", *106th European Signal Processing Conference*, 2008

[17] International Telecommunication Union, "Low-complexity coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss", ITU-T Recommendation G.722.1 Annex C, Geneva, Switzerland, 2005

[18] International Telecommunication Union, "Low-complexity fullband audio coding for high-quality conversational applications", ITU-T Recommendation G.719, Geneva, Switzerland, 2008

[19] International Telecommunication Union, "Method for the subjective assessment of intermediate sound quality (MUSHRA)", ITU-R, Recommendation BS. 1543-1, Geneva, Switzerland, 2001